LATTICE VECTOR QUANTIZATION FOR THE ANALYSIS OF MOLECULAR DATA

Cédric RAMASSAMY[1,*], Vincent RICORDEL[1], Oana CRAMARIUC[2], Bogdan CRAMARIUC[3]

[1]LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597 (Institut de Recherche en Communications et Cybernétique de Nantes), Polytech Nantes, rue Christian Pauc BP 50609 44306 Nantes Cedex 3; [2]Department of Physics, Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland; [3]IT Center for Science and Technology, Av. Radu Beller 25, Bucharest, Romania.

cedric.ramassamy@univ-nantes.fr, vincent.ricordel@univ-nantes.fr, oana.cramariuc@tut.fi, bogdan.cramariuc@citst.ro

ABSTRACT

We introduce a novel simplification scheme of molecular data using Tree-Structured Lattice Vector Quantization (TSLVQ). The method, based on the embedding of truncated lattices, permits also hierarchical description of the 3-D volume through a tree-structure. We apply TSLVQ to simplify the electron densities maps of three thrombin inhibitors (MQPA, NAPAP, 4-TAPAP).

Keywords: *computational chemistry, electron density, simplification, tree-structure, lattice, vector quantization, multiresolution, molecular similarity, MQPA, NAPAP, 4-TAPAP*.

1. INTRODUCTION

During the last decades signal processing methods have been employed beyond their traditional application domains and have started to contribute significantly to advancements in biology, biochemistry and biomedicine. What was earlier viewed as digital signal processing represents nowadays only a small part of the new concept of signal processing which can be described as the collection of methods for analyzing, manipulating and presenting natural information. One underlying cause of this expansion is the exponentially growing volume of numerical data obtained through modeling and simulation techniques or by employing modern high-throughput experimental investigation tools such as DNA-, protein-, cellular- and antibody-microarrays. One other cause is the substantial effort employed in developing systemic models of processes taking place in living organisms, models which combine individual processes into a larger coherent picture.

The computing power has now increased to a level high enough to process large amounts of molecular data. Nowadays large systems of 10.000 – 50.000 atoms are approached computationally. Beyond the mere resources limits of processing such a huge amount of data, the major task is to detect and extract semantic knowledge in the molecular system. Thus, more than data compression, a simplification of the molecular topology can help in this task. Additionally, Critical points (CP) and singularities, computed from the simplified molecular data, become representatives of groups of atoms rather than atoms themselves [1].

With specific molecular properties conferred by functional groups (e.g. hydroxyl –OH and carboxyl -COOH) efficient ways to identify these groups are sought [1]. This process can lead to simplified representations of the molecules, for e.g. amino acid residues in proteins can be depicted using a lower level representation, *i.e.* two or three pseudo-atoms representing the whole residue. Such representations are of importance for drug design applications and to overcome structural inaccuracies issued from experimental data such as X-ray analysis. In protein studies, for instance, researches focus on the detailed characterization of macromolecular surfaces and topologies [2]. Synthesis of drug-like molecules is driven by the 3D structural characteristics of molecules [3].

Previous works on simplification focus on a crystallography-based formalism, whereas a recent trend aims at adapting methods taken from signal processing domain. Exactly, the electron densities are decomposed into successive resolution of wavelets. CP analysis on low wavelet resolution showed relevant results [1]. In addition, from the image processing domain, Vector Quantization (VQ) [4] was adapted to electron density simplification, in particular a neural network method based on a cost function [2]. The technique of VQ represents the data into a set of reproduction vectors or codewords while keeping the sensible shape of the source density map.

In the present work we employ Vector Quantization, a method typically used in transmission (source coding) and data classification, as a first step in processing of molecular electron densities for further molecular similarity analysis, topological exploration and visualization. The considered electron densities are obtained from quantum mechanical calculations which, as opposed to previous attempts, allow for a full control of the input data quality and a good understanding of the structure-density relationship. The paper is organized as follows: In the first part the Tree-Structured Lattice VQ **(**TSLVQ) approach is detailed, the second part deals with the simplification of the Electron Density Maps (EDM) using TSLVQ, results are described in the third part and a conclusion is given at the end.

2. QUANTIZATION

2.1. Vector Quantization

A concise description defines quantization as the computation of an approximation for a given signal [4]. VQ consists in representing, an ordered set of numbers (or vectors) by a more reduced set called the codebook [4,5]. VQ is mainly

applied to source encoding in order to shrink the volume of data needed to represent the information [5]. Our method deals with decreasing the granularity of the source representation in order to characterize hierarchically the information.

2.2. Lattice Vector Quantization

2.2.1. General

The goal of Lattice VQ (LVQ) is to strongly structure the codebook in order to reduce the computing complexity involved in the codebook design [5]. LVQ does not require a training step nor an exhaustive search in order to build the codebook. The codewords are points of a truncated lattice regularly scattered into space [6]. The codewords are also the centres of the *Voronoi* cells, the duals of lattices. The space embedding of the lattices will determine their properties. Unlike in a LBG-based (Linde, Buzo, Gray algorithm) method [4,7], in LVQ the encoding is related to the coordinates of the vector. Thus, it is only based on rounding and scaling operations. Therefore, the coding is very simple [5,6].

2.2.2. Lattices

In [6] it is shown that all lattices have not necessarily an optimal space embedding. Moreover, the efficiency of the quantization is directly linked to the geometry of the chosen polytope which describes the *Voronoï* cell. The choice of the best lattice in the molecular situation leads to an orientation independence feature. The EDMs are three-dimensional matrices, thus we need 3-D lattices.

There are only three lattices for dimension 3 for which fast quantization algorithms are known: **Z3** Cubic lattice, **D3** the Rhombic dodecahedron and **D3*** the Truncated Octahedron. Figure 1 shows a 2-D lattice embedding example with the hexagon.

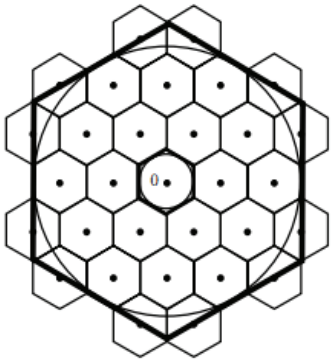

Figure 1: *A sub optimal embedding - the hexagonal Lattice* [8].

2.3. Tree-Structured Vector Quantization

Tree Structured VQ is a gathering of many quantization approaches where the quantization is processed through a decision tree [4]. Its advantages are reduced computation with the use of simpler sub-codebooks, and a structure adapted to progressive representation.

2.4. Tree-Structured Lattice Vector Quantization

Tree Structured LVQ (TSLVQ) [9] aims at using a hierarchical set of embedded lattices which is achieved such as it is possible to embed a lower scale truncated lattice into a cell of the next higher scale truncated lattice. So a scaling factor between consecutive lattices of the hierarchy has to be set (Figure 2). The principles of the TSLVQ are [8]:

1. A source vector is projected into a first truncated lattice;
2. To get a finer quantization, another lower scale truncated lattice is embedded into the *Voronoï* cell where lies the input vector;
3. The previous operation can be repeated.

It is more convenient to deal with the input vector scale than to use several lattices with different scales. The principles of the encoding are shown at Figure 3.

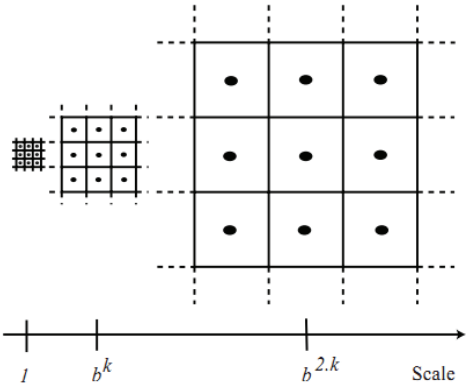

Figure 2: *Hierarchical set corresponding to the cubic lattice. Here, the scaling factor b=3* [8]**.**

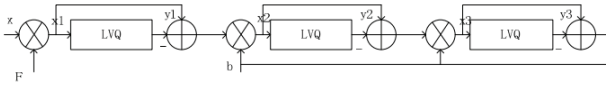

Figure 3: *The principle of TSLVQ* [8]*: x is the source vector, y_n the successive reproduction vector, F the first scaling factor and b the next scaling factors.*

Where we have: the scaling factor F used to project the input vector *x* into the first truncated lattice:

$$F = \frac{b \times \rho}{L_{2\max}}$$

where ρ is the corresponding packing radius, and $L_{2\max}$ the maximal L_2 norm of *x*. So, all the inputs are projected into a hyper-sphere whose radius equals $b \times \rho$.

In the normalized space, the scaling factor used to project each translated vector into the next truncated lattice of the hierarchy is *b*. The reproduction vector of the truncated lattice for the j^{th} level is y_j. The final value of the reproduction vector associated with x will be then:

$$y = \frac{1}{F} \times \sum_j \frac{y_j}{b^{j-1}}$$

with *j* the level of the quantization. At each step, the same fast quantization algorithm is used.

We introduce a novel simplification scheme of molecular data based on TSLVQ. This kind of VQ is very fast thanks to the LVQ and its tree structured codebook. It is, therefore, well adapted to the analysis of the distribution of the source.

3. METHODS

The data source is a *Gaussian cube* file [10], the output data are a set of *cube* files, from the TSLVQ step. Output *cube* files are snapshots of the different levels of quantization (Figure 4).

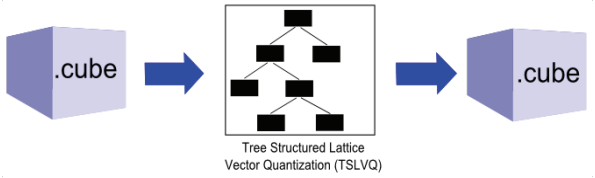

Figure 4: *General scheme of the method.*

3.1. Electron densities

The electron density of a chemical entity plays a critical role in explaining and understanding its properties. In fact, it was proven in 1964 by Hohenberg and Kohn that the electron density determines all properties of the ground state of a chemical system, including the ground state energy. Accordingly, the work of Popelier, based on Bader's Atoms in Molecules theory, proves that it is relevant to exploit 3D electron density data by taking advantage of its topology [11,12]. It has also been shown that the topological analysis of electron density allows simplification of the 3D distribution in reduced representations without losing significant information. In this work, we employ quantum mechanical calculations at the density functional theory level to calculate the EDMs of the compounds described in section 4. The Per-dew-Burke-Ernzerhof exchange-correlation functional to-gether with a TZVP basis set were used as implemented in the Gaussian 09 computational package *www.gaussian.com*.

3.2. Vector Quantization of the electron densities.

3.2.1. Data loading

A Gaussian *cube* file describes the 3-D maps of EDMs or of electrostatic potentials but the present work focuses on EDMs only. The latter stands for a grid that slices the space of the studied molecule. Three data sets of *Gaussian cube* [10] files were used with the new TSLVQ method. We focused our study on three anticoagulants, thrombin inhibitors (MQPA, NAPAP, 4-TAPAP), which were previously used by Leherte as test systems for the wavelet based simplification [1]. Each of the molecules is described in three resolutions of electron densities through their dedicated *cube* file. Given the amount of data, the EDMs need to be batch processed. Flaws of the data slicing should introduce sets of non-trivial errors. Indeed, the size of the batch directly impacts the precision of the quantization.

3.2.2. Quantization with TSLVQ

TSLVQ with *cube* files source needs two parameters: scaling (or boxing) factor and quantization level. There are two ways of constructing the tree: a) "Merging" way consists in building the whole tree in a first step. Then the tree has to be fully browsed in order to merge leaves successively. b) "Splitting" way constructs the tree progressively at each iteration of the quantization. At the first level all points are combined in one single node, this is the most degraded representation of the source, and at the next levels, if the processing rule requires it; a leaf is splitted in a range of children leaves and so on. Using a higher number of quantization levels, we get a more detailed (and voluminous) description of the molecule.

For obvious reasons of resources saving, we have implemented the splitting way. The coordinates of the density are rounded to be aligned on the current lattice reproduction vectors. Once all data are processed a mean is calculated between all the densities of the current codeword. The scaling factor controls the number of possible children for the current node (e.g.: for the cubic lattice, if b=3 in 3-D → 3^3= 27 children possible for each node but only those whose dual *Voronoï* cell contains densities are created). The existing lattice is rescaled for the current child node, and all the densities that belong to the parent node are quantized. The process stops when it reaches the quantization level given in parameter. The quantization tree resulting is the input of the next step.

3.3. Visualization and analysis of the results

The method produces as result a set of *cube* files (one file per level). This allows the visualization and analysis with common computational chemistry tools (e.g.: *Gaussian* [10]).

4. RESULTS

The compounds have a star shape with three branches completed by a cyclic substructure bound to sulfonyl function, piperidine ring, or amidino group (Figure 5).

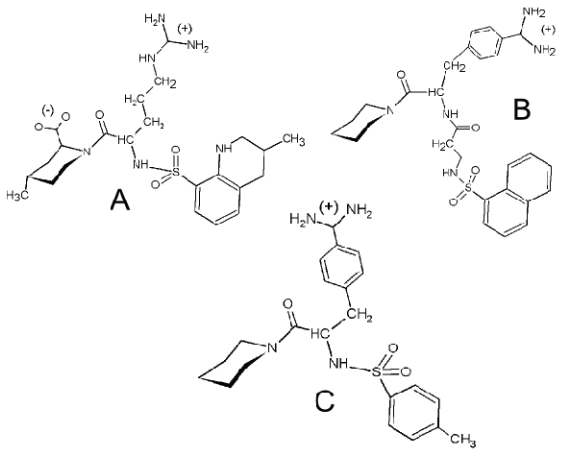

Figure 5: *2-D structures of compounds MQPA (A), NAPAP (B), and 4-TAPAP (C)*

In Figures 6-8 the EDMs are visualized using a blue (low)-to-red (high) colour map and a transparency coefficient proportional to the voxel value. When analyzing Figures 6-8, it is important to keep in mind that the EDMs obtained by TSLVQ have a level of detail proportional to the number of quantization steps. Subsequently, with 5 levels of quantization, the obtained EDMs are closest to the source data. In each case the oxygen atoms of the sulfony, carbonyl and/or carboxy groups lead to peaks in the EDMs obtained after 5 levels of quantization. Such groups should thus easily match together in a superposition algorithm using a density based

similarity measure. Other atoms in the functional groups like for e.g. nitrogen and sulfur (see Figure 5) are not detected at this resolution.

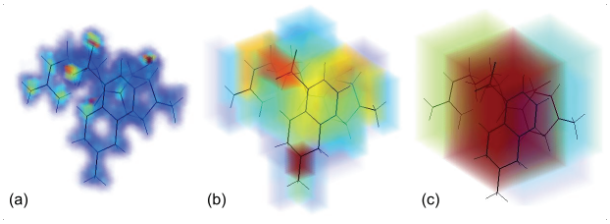

Figure 6: *3-D MQPA EDM (resolution 113×96×93) with 5, 3 and 2 levels of TSLVQ (Z3, b2), respectively.*

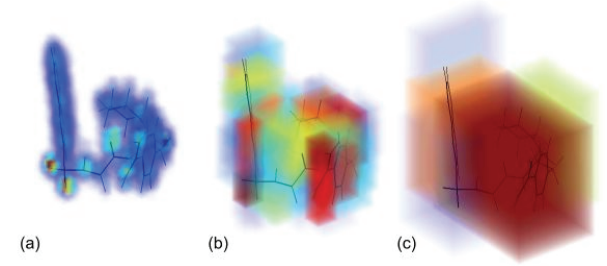

Figure 7: *3-D NAPAP EDM (resolution 120×98×85) with 5, 3 and 2 levels of TSLVQ (Z3, b2), respectively.*

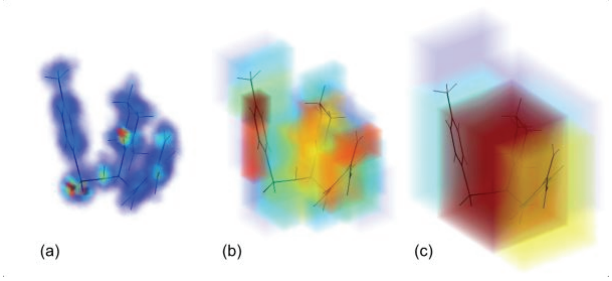

Figure 8: *3-D 4-TAPAP EDM (resolution 119×94×89) with 5, 3 and 2 levels of TSLVQ (Z3, b2), respectively.*

The terminal amidino groups become visible in the third quantization level in which they bear clear EDM peaks. At the same time also the other nitrogen atoms like the ones in the hetero-cycles become visible. Independent on the quantization level no peaks are identifiable with the alkyl part of MQPA implying that chemical groups composed of only C atoms can remain undetected. We do expect however that aromatic rings connected to functional groups to produce clear peaks in the EMS due to the disruption in π-conjugation. Further studies are needed to justify this assumption. The order of the peak density values can be approximately summarized as: sulfonyl ≥ (CO(O) > amidino/N > alkyl chain.

The above analysis, although performed at a qualitative level, agrees well with the results obtained from employing wavelet quantization [1]. Some differences exist though like for e.g. the identification of the sulfonyl groups which is done in our case based on the peaks created by the oxygen atoms rather than by sulfur. This is why in our case sulfonyl groups become less distinguishable from other oxygen bearing functional groups like carboxyl. Nevertheless, we can conclude that the same atoms or functional groups can be identified at certain quantization levels of all molecules considered in this study which could in principle make the theoretical assignment of a chemical group to a peak possible. However, this assertion needs to be validated through further work implementing for e.g. critical points analysis and graph representation of the electron density topology.

5. CONCLUSION

In this paper we showed that TSLVQ is well adapted to the simplification of EDMs. Using a cubic lattice we obtain a set of descriptions hierarchically organized in levels (with one cube file per level): the deeper the level the finer the resolution. Result are presented for three molecules (MQPA, NAPAP, 4-TAPAP) and compared qualitatively with the ones obtained by Leherte [1]. The next step will be to include the computation of critical points directly into the 3-D hierarchical description (the tree-structure).

REFERENCES

[1] L. Leherte, "Application of multiresolution analyses to electron density maps of small molecules: Critical point representations for molecular superposition," *Journal of Mathematical Chemistry*, vol. 29, 2001, pp. 47-83.

[2] J. Burton, N. Meurice, L. Leherte, and D.P. Vercauteren, "Can Descriptors of the Electron Density Distribution Help To Distinguish Functional Groups?," *Journal of Chemical Information and Modeling*, vol. 48, 2008, pp. 1974-1983.

[3] P.A. De-Alarcón, A. Pascual-Montano, A. Gupta, and J.M. Carazo, "Modeling shape and topology of low-resolution density maps of biological macromolecules.," *Biophys J*, vol. 83, 2002, pp. 619-632.

[4] A. Gersho and R.M. Gray, *Vector quantization and signal compression*, Norwell, MA, USA: Kluwer Academic Publishers, 1992.

[5] V. Ricordel, "Etude d un schéma de quantification vectorielle algébrique et arborescente. Application à la compression de séquences d'images numériques, PhD thesis," Université Rennes 1, 1996.

[6] J.H. Conway, N.J.A. Sloane, and E. Bannai, *Sphere-packings, lattices, and groups*, New York, NY, USA: Springer-Verlag New York, Inc., 1993.

[7] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, 1980, pp. 84-95.

[8] V. Ricordel and C. Labit, "Vector Quantization by packing of embedded truncated lattices," *IEEE International Conference on Image Processing (ICIP)*, Washington DC: 1995.

[9] V. Ricordel and C. Labit, "Tree-Structured Lattice Vector Quantization," *European Signal Processing Conference (EUSIPCO)*, Trieste, Italy: 1996.

[10] P. Bourke, "Gaussian Cube Files specifications - http://paulbourke.net/dataformats/cube/," 2003.

[11] P.L.A. Popelier, "Molecular similarity and complementarity based on the theory of atoms in molecules," *Molecular similarity in drug design, Dean, P. M.*, 1995, pp. 215-240.

[12] P.L.A. Popelier, "Integration of atoms in molecules: A critical examination.," *Mol. Phys*, vol. 87, 1996, p. 1169–1187.